

非广延熵:一种新的 Internet 拓扑度量

张益方¹ 钱亚冠¹ 王 滨² 吕兵兵¹

(浙江科技学院理学院 杭州 310023)¹ (浙江大学计算机学院 杭州 310027)²

摘 要 随着 Internet 规模的不断扩大,Internet 拓扑模型已成为网络领域的研究热点。从最初的 Waxman 随机模型到基于幂律^[1]特征提出的 BA 模型、GLP 模型、PFP 模型,都试图去发现和解析 Internet 拓扑演化的规律。节点度、聚类系数、介数等拓扑度量分别被提出,用于刻画拓扑特征。引入统计物理学上非广延熵^[2]作为一个新的拓扑度量,从概率分布的均匀程度上刻画拓扑模型的节点度分布的特征。为了定量地分析非广延熵在不同模型上的差异,通过从 CAIDA 获得最新的真实拓扑数据,比较 BA 模型和 Waxman 模型与真实拓扑数据在非广延熵上是否具有一致性,来验证非广延熵作为拓扑度量的合理性。最后指出 BA 模型的非广延熵接近真实的 Internet 拓扑,从而也说明非广延熵作为一种新的拓扑度量的可行性。

关键词 Internet, 拓扑模型, 非广延熵

中图法分类号 TP393.01

文献标识码 A

Non-extensive Entropy: A New Metric of Internet Topology

ZHANG Yi-fang¹ QIAN Ya-guan¹ WANG Bin² LU Bing-bing¹

(Science College, Zhejiang University of Science and Technology, Hangzhou 310023, China)¹

(College of Computer Science, Zhejiang University, Hangzhou 310027, China)²

Abstract With the size of Internet growing fast, the topology model of Internet is becoming an appealing point in research community. People make try to put deep insight into the nature of Internet topology, by constructing more models from initial Waxman random model to BA, GLP, and PFP. A number of topology metrics were proposed, including node degree, cluster, and betweenness, to describe the topology characters. Non-extensive entropy, which is employed extensively in statistical physics, was introduced for a new Internet topology metric in this paper. It is used to depict the purity of node degree distribution. In order to evaluate the non-extensive entropy as topology metric, we analyzed the consistency of BA, Waxman and real topology generated from CAIDA, by their non-extensive entropies. Finally, we argued that BA is more in accordance with real Internet topology by comparing their non-extensive entropy, which confirms that non-extensive is fit for topology metric.

Keywords Internet, Topology model, Non-extensive entropy

1 引言

Internet 是复杂网络 (complex network) 的一个典型例子,也可以说是一个庞大的人工生态系统。Internet 发展到现在,几乎没有人说得清楚它目前的真实拓扑结构是什么样的。对 Internet 的拓扑结构的研究,前人已经做了很多工作,从最初的随机图模型到结构模型,再到 Faloutsos 提出的节点度的分布幂律特征^[3,4]、Barabasi 等提出的无标度网络,以及后来的 Zhou S 等提出的 rich-club 现象、Mahadevan 等关于节点度的相关性研究,可以说都取得了不少成果。

对 Internet 拓扑模型的研究意义在于:①深入地理解现存路由协议的局限性,正确地评估新的协议、体系的设计,预测将来互联网的发展需求;②深入地理解网络技术、拓扑结构和背后的经济因素之间的相互作用关系;③路由协议的模拟

与仿真、网络病毒的传播等均需要网络拓扑模型;④发现拓扑演变背后的动因、本质规律,对于指导和预测将来在 Internet 上部署各种新技术、新协议和新应用将具有重要指导意义。

真实的 Internet 拓扑极其复杂,而拓扑模型是 Internet 的一个高度的数学抽象,而对它的建模不能按“还原论”观点进行完全复制重现,需要提取它影响协议性能、部署等方面的关键特征,去发现它的特征度量。这些特征度量不仅有表征局部的,也有表征全局结构的。评价拓扑模型的优劣,不能仅以单个的特征度量为依据。如何正确评价 Internet 拓扑模型仍然是一个困难的公开性问题。

2 拓扑模型概述

2.1 Waxman 随机模型

Waxman 模型是随机图拓扑模型中的典型代表,该模型

本文受国家科技支撑计划课题(2008BAH37B02),国家 973 计划重大研究课题(2007CB307102),浙江省大学生科技创新活动计划(新苗人才计划 2010R415)资助。

张益方(1989—),男,本科,主要研究方向为 Internet 拓扑建模,E-mail:qianyg@yeah.net;钱亚冠(1976—),男,博士生,讲师,主要研究方向为下一代互联网;王 滨(1976—),男,博士,副教授,主要研究方向为下一代互联网;吕兵兵(1989—),男,本科,主要研究方向为 Internet 拓扑建模。

在1988年由Waxman提出。Waxman模型以任意两节点间的距离为自变量来计算两点间直接相连的概率,其概率函数为 $P(\mu, \nu) = \alpha e^{-d/(\beta L)}$,其中 $\alpha > 0, \beta \leq 1; d$ 为 μ 到 ν 的欧氏距离; L 为两点间的最长距离。当增加 α 时,所建模型将有更多的短边、更长的跳数直径、更短的长度直径;增加 β 将增加模型中长边所占的比例。Waxman模型的平均节点度为 $E[n\alpha e^{-d/(\beta L)}] = n\alpha E[e^{-d/(\beta L)}]$ 。对于Waxman模型的目标边数,有一组 α, β 集合可以保证达到。当模型参数 α, β 固定时,参数 L 对模型的边数几乎没有影响,因为虽然 L 的改变会影响两节点间距离 d 的值,但是 d/L 的值保持不变。

由于在基于随机图的拓扑模型中,节点度数也会随着节点数量的增加而增加,因此,随机图拓扑模型无法生成节点众多且节点平均度较小的网络。

2.2 Barabasi-Albert 模型^[6,7]

Barabasi和Albert于1999年提出,大型网络能自组织成无标度(scale-free)状态的原因是:

- ①增长(growth)特性,即网络的规模是不断扩大的;
- ②优先连接(preferential-attachment)特性,即新的节点加入网络,总是倾向于与较高连接度的“大”节点连接。

BA模型就是基于上述两个特征构造的具有度分布呈幂律特征的无标度网络。构造算法如下:

- ①增长:从一个具有 m_0 个节点的网络开始,每次加入一个新的节点,并连接到 m 个已存在的节点上,且 $m \leq m_0$;
- ②优先连接:新加入的节点以概率 $\Pi(k_i) = k_i / \sum_j k_j$ 与节点 i 连接, k_i 是节点 i 的度。

3 拓朴度量

所谓的拓朴度量就是要能反映出拓朴模型的本质特征,一般通过统计方法得出。Internet经过几十年的演化,目前节点度的分布广泛接受遵循幂律分布的规律。

3.1 节点度分布(degree distribution)

节点度分布是拓朴的一个统计特征。按照图论的定义,节点的度就是指与该节点有直接边相连的节点的个数。节点度反映了该节点的连接能力,也是作为其他统计指标的一个基本分类依据。

3.2 度的相关性(degree correlation)

度的相关性也称度的联合分布(JDD),它表征拓朴的结构特征。节点度分布只表明了在给定度的情况下,有多少节点满足条件。它不能反映节点之间的互连情况,而节点度的相关性用于反映节点与其邻接点的关系。

3.4 聚簇(Clustering)

节点的度的联合分布只给出了相邻节点间关于度的信息,并没有给出邻接点之间是如何互连的。聚簇给出了相邻节点构成小集团的紧密程度的衡量指标。聚簇系数描述了网络中节点的邻接点之间也互为邻接关系的比例,反映小集团结构的完美程度。

3.5 介数(betweenness)

介数分为节点介数和边介数。一个节点的介数指经过该节点的最短路径数目。该节点的介数值越大,表明该节点在网络中的枢纽地位越高。一条边的介数指通过该条边的所有最短路径的数目,同样也反映了这条边的枢纽地位。

4 非广延熵

熵的概念最初作为一个热力学概念,经过波尔兹曼、吉布斯等人的努力,成为统计物理学的重要基础。后经香农的拓展,将熵的概念应用到信息领域。从热力学的角度看,熵是系统微观态分布均匀性的一种度量,分布越均匀,熵越大,反之则越小^[8]。根据熵的这一特性,我们可以用它来度量Internet这样的复杂网络。在Internet网络拓扑中,节点度的分布越均匀,熵值越大,反之越小。

下面我们引入Tsallis非广延熵的概念,它由巴西物理学家C. Tsallis提出。非广延性是指下面的赝可加性:

$$\frac{S_q(A+B)}{k} = \frac{S_q(A)}{k} + \frac{S_q(B)}{k} + (1-q) \frac{S_q(A)}{k} \cdot \frac{S_q(B)}{k} \quad (1)$$

式中, $A+B$ 表示当前系统由 A, B 两个子系统组成。 q 值表明了系统的非广延程度。

非广延熵克服了波尔兹曼-吉布斯熵的局限性,适合于描述具有长相关、长记忆的系统。而Internet的拓朴被证明具有长相关幂律分布特征。为此我们重新定义适合表达Internet拓朴的非广延熵,来定量地度量Internet拓朴模型。

定义1 我们称:

$$S_q = \frac{1 - \sum_{i=1}^N P_i^q}{q-1} \quad (q \in R^+) \quad (2)$$

为Internet拓朴的非广延熵,其中, N 为网络节点数, q 为非广延参数, $\{P_i\}$ 为节点度的概率集合,其满足归一化条件: $\sum_{i=1}^N P_i = 1$;当 $q < 1$ 时,称其为超广延;当 $q > 1$ 时,称其为次广延;当 $q = 1$ 时,称其为广延,数值上等于Shannon熵:

$$S = - \sum_{i=1}^N P_i \ln(P_i) \quad (3)$$

定义2 我们称:

$$S = \begin{cases} \frac{1 - \sum_{i=1}^N P_i^q}{q-1} / \frac{1 - \sum_{i=1}^N (1/N)^q}{q-1}, & (q \in R^+ \cap q \neq 1) \\ - \sum_{i=1}^N P_i \ln(P_i) / \log(N), & (q = 1) \end{cases} \quad (4)$$

为标准熵,其中, N 为网络节点数, q 为非广延参数, $\{P_i\}$ 为节点度的概率集合,满足归一化条件: $\sum_{i=1}^N P_i = 1$ 。

有了对Internet拓朴的非广延熵定义,就可以对上述两种Internet拓朴模型进行分析,需要强调的是这里的非广延熵是对网络节点度分布的一种度量,其熵值与拓朴模型算法紧密相联,这些算法决定了网络中节点与节点之间的优先连接能力,从而生成了不同的节点度分布,即式(2)中的 $\{P_i\}$ 。式(4)是为了衡量模型在不同节点总数下的拓朴度分布,因此专门引入标准化的非广延熵,实验中将标准化后的非广延熵进行比较。

5 实验分析

为了验证拓朴模型,从CAIDA(网络测量权威机构)获得了最新的Internet拓朴数据。通过比较在不同节点规模的拓朴下节点度的非广延熵,试图发现非广延熵作为一种新的度量标准的合理性。在实验中,分别选择了非广延参数 $q = 0.8, q = 1, q = 1.2$ 来对Waxman、BA模型以及CAIDA拓朴数据进行模拟,得到图1—图3的结果。

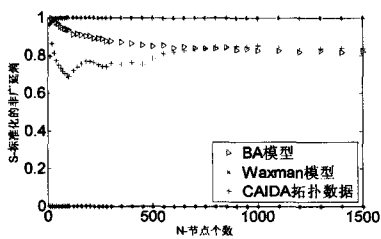


图1 非广延参数 $q=0.8$

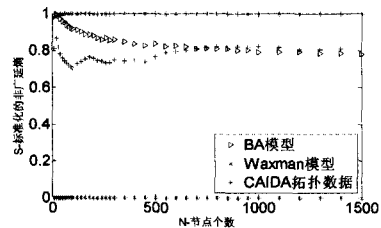


图2 非广延参数 $q=1$

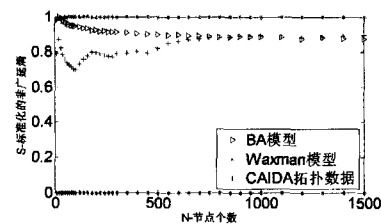


图3 非广延参数 $q=1.2$

从图中可以看出, Waxman 模型的标准熵趋于 1, 说明随机程度越大, 节点度分布也越均匀。BA 模型及 CAIDA 拓扑数据得到的标准熵逐步趋于稳定, 而且在确定的非广延参数 q 下, N 达到一定数量后, 两者的稳定状态十分相近。

为了更好地分析这一现象, 对不同 q 值在 1500 个节点下进行了不同模型的非广延熵值的计算, 得到表 1。

表 1 不同 q 值下的非广延熵

非广延参数 q	CAIDA	BA	Waxman
0.2	0.9078	0.8951	0.9993
0.4	0.8440	0.8203	0.9989
0.6	0.8127	0.7721	0.9989
0.8	0.8168	0.7815	0.9991
1.0	0.8426	0.8223	0.9994
1.2	0.8971	0.8756	0.9997
1.4	0.9382	0.9194	0.9999
1.6	0.9660	0.9600	0.9999
1.8	0.9821	0.9807	1.0000

对表格数据进行拟合得到图 4。

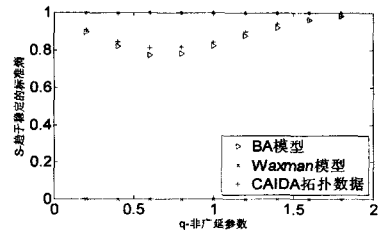


图4 非广延参数与标准熵的关系

从图 4 中可以看出, 在相应的广延参数下得到的稳态、BA 模型及 CAIDA 拓扑标准熵值十分相近。实验结果表明, BA 模型与当前的网络拓扑结构十分相近, 对网络拓扑结构的研究可以借用 BA 模型。这从另一个侧面证明了非广延熵适合作为一种新的拓扑度量。但 Waxman 模型与真实网络拓扑还是存在一定的差距。

结束语 本文引入了非广延熵、信息熵及标准熵的定义, 通过对 BA 模型、Waxman 模型以及真实拓扑数据进行模拟, 用标准化的非广延熵定量地刻画, 得出 BA 模型是符合真实网络结构的, 基于 BA 模型对网络拓扑的研究是合理的。理论模型是理想化的, 在真实的网络中存在路由节点之间的断开与重连操作, 这是一个动态的过程, 而理论模型忽略了这些因素。为了进一步地模拟真实的网络, 下一段的工作需要对理论模型做相应的改进, 考虑增加影响网络结构的因素。

参考文献

- [1] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M]. 北京: 清华大学出版社, 2006: 210-232
- [2] 曹克非, 王参军. Tsallis 熵与非广延统计力学[J]. 云南大学学报: 自然科学版, 2005, 27(6): 514-520
- [3] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology[C]//Proc. of ACM SIGCOMM. 1999
- [4] Tian Bu, Towsley D. On Distinguishing between Internet Power Law Topology Generators[C]//Proceeding of INFOCOM. New York, 2002, 2: 638-647
- [5] 叶中行. 信息论基础(第二版)[M]. 北京: 高等教育出版社, 2007: 32-56
- [6] Barabasi A, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286: 509
- [7] Zhou Shi, Mondragon R J. Accurately modeling the Internet topology[J]. Physical Review E, 2008, 70(6): 66-108
- [8] 谭跃进, 吴俊. 网络结构熵及其在非标度网络中的应用[J]. 系统工程理论与实践, 2004(6): 32-36

(上接第 139 页)

- [7] Nallapati R. Semantic Language Models for Topic Detection and Tracking[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 Student Research Workshop. Edmonton, Canada, Association for Computational Linguistics, 2003: 1-6
- [8] Connell M, Feng A, Kumaran G, et al. UMass at TDT 2004[C]//Proceedings of the TDT2004 Workshop. 2004
- [9] Nigam K, McCallum A, Thrun S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000, 39(2/3): 103-134
- [10] Hofmann T. Probabilistic latent semantic indexing [C]// Proceedings of the Twenty-Second Annual International SIGIR Conference. 1999

- [11] Boyd-Graber J L, Blei D M. Syntactic Topic Models[C]//NIPS 2008. 2008: 185-192
- [12] Blei D M, Lafferty J D. Correlated Topic Models [C]// NIPS 2005. 2005
- [13] Blei D M, Lafferty J D. Dynamic topic models[C]//ICML 2006. 2006: 113-120
- [14] 曹娟, 张勇东, 李锦涛, 等. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报, 2008, 31(10): 1780-1787
- [15] 王会珍, 朱靖波, 季铎, 等. 基于多向量模型的中文话题追踪: 自然语言理解与大规模内容计算[D]. 南京: 清华大学出版社, 2005: 669-671